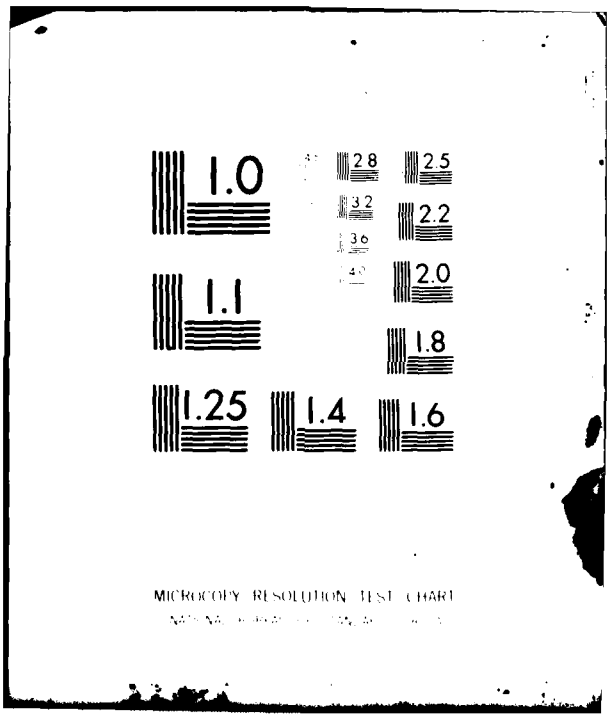
AD-A109 660 STANFORD UNIV CA DEPT OF STATISTICS F/G 12/1
ON THE ACCURACY OF SIMULATED PERCENTAGE POINTS.(U)
OCT 81 J M JURITZ, J W JURITZ, M A STEPHENS N00014-76-C-0475
UNCLASSIFIED TR-309 NL

END
DATE FILMED
02-82
DTIC


1.0
2.8
2.5
3.2
2.2
3.6
2.0
1.1
1.8
1.25
1.4
1.6
MICROCOPY RESOLUTION TEST CHART

AD A109660

ON THE ACCURACY OF SIMULATED PERCENTAGE POINTS

By

J.M. Juritz, J.W.F. Juritz and M.A. Stephens

TECHNICAL REPORT NO. 309

OCTOBER 7, 1981

Prepared Under Contract
N00014-76-C-0475 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted
for any Purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist
A

ON THE ACCURACY OF SIMULATED PERCENTAGE POINTS

By

J.M. Juritz, J.W.F. Juritz, and M.A. Stephens

Abstract.

The estimates of percentage points of a distribution, obtained from one large sample of Monte Carlo values, are compared with those obtained by dividing the large sample into several subsamples and taking the average. The single-sample method appears to be preferable.

Key Words. Monte Carlo methods; percentage points; significance points; simulated percentage points; simulations.

Submitted for publication to: JASA

ON THE ACCURACY OF SIMULATED PERCENTAGE POINTS

By

J.M. Juritz, J.W.F. Juritz and M.A. Stephens
University of Cape Town and Simon Fraser University

1. Introduction.

In recent years Monte Carlo simulation methods have been used more and more frequently to produce tables of percentage points of probability distributions when mathematical difficulties prevent an analytic solution to the problem. Schafer (1974) has outlined possible sources of error and presented a method for calculating a confidence interval for a percentile. This is based on several Monte Carlo runs. This note is a revision of a previous manuscript (Juritz, Juritz, and Stephens, 1978). In it we present an alternative method based on the order statistics in one run, which has certain advantages which we discuss below. We illustrate the two methods with simulations for percentage points of the normal distribution.

2. Confidence Intervals Using Order Statistics.

Let $f(x)$ be the probability density function of a continuous random variable X and let ξ_p denote the p^{th} percentile of X. Suppose a random sample of size n is drawn from $f(x)$ and let $X_{(i)}$ be the i^{th} order statistic, i.e. the i^{th} smallest observation. David (1970) shows that for any continuous random variable the random interval $(X_{(r)}, X_{(s)})$ with $r < s$ covers ξ_p with probability $\pi(r,s,n,p)$ given by

$$\text{Prob}(X_{(r)} \leq \xi_p \leq X_{(s)}) = \pi(r,s,n,p) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} . \tag{2.1}$$

This probability does not depend on $f(x)$, and so is very suitable for use in simulation studies, where $f(x)$ is always unknown. In this application, n is the number of repetitions, $p = \text{Prob}(X \leq \xi_p)$ and r and s must be chosen so that the interval $(X_{(r)}, X_{(s)})$ has the desired confidence level $1-\alpha$; i.e. we choose r and s so that $\pi(r,s,n,p) = 1-\alpha$.

Since in a simulation n is very large indeed the value of $\pi(r,s,n,p)$ can be evaluated using the normal approximation to the binomial distribution, and r and s can then be found. Let $z_{1-\alpha/2}$ be the $100(1-\alpha/2)^{\text{th}}$ percentile of the standard normal distribution. Then

$$r = -z_{1-\alpha/2}(np(1-p))^{1/2} + np + 1/2 \quad \text{and}$$

(2.2)

$$s = z_{1-\alpha/2}(np(1-p))^{1/2} + np + 1/2 \ .$$

The expressions for both r and s end in $(+ 1/2)$ because the sum in (2.1) runs from r to $s-1$, not to s. Let $k = [np]+1$, where $[np]$ denotes the largest integer less than or equal to np, and define

(2.3) $$X_p = X_{(k)} \ ;$$

X_p is the usual estimate of ξ_p and the interval $L \equiv (X_{(r)}, X_{(s)})$ gives the confidence interval for ξ_p, at confidence level $100(1-\alpha)\%$, since it covers ξ_p with probability $1-\alpha$ (approximately, but to high accuracy). We shall refer to the above as the single-sample method. Schafer (1974)

proposes an alternative method based on the fact that X_p is asymptotically normally distributed with mean ξ_p and variance

$$\sigma^2 = \left(\frac{p(1-p)}{n}\right)\left(f(\xi_p)\right)^{-2} .$$

Instead of a single run with n repetitions, c runs each with m repetitions $(n = cm)$ are made. For each run the sample percentile $X_{([mp]+1)}$ estimates ξ_p. Denoting these estimates by $X_{p1}, X_{p2}, \ldots, X_{pc}$, the percentile ξ_p is estimated by the average

$$\bar{X}_p = \frac{1}{c} \sum_{i=1}^{c} X_{pi} \tag{2.4}$$

and the sample variance

$$s^2 = \sum_{i=1}^{c} (X_{pi} - \bar{X}_p)^2/(c-1) \tag{2.5}$$

estimates σ^2. Then

$$(\bar{X}_p - t_{1-\alpha/2}\, sc^{-1/2},\ \bar{X}_p + t_{1-\alpha/2}\, sc^{-1/2})$$

is a confidence interval for ξ_p with confidence probability $1-\alpha$, where $t_{1-\alpha/2}$ is the upper $100(1-\alpha/2)^{th}$ percentile of Student's t with $c-1$ degrees of freedom. We shall call this the multi-sample method. The number of runs must be chosen to give a good estimate of the standard deviation.

3. Comparison of the Two Methods.

3.1. Value of the estimated percentage point.

Since in practice everyone who does a Monte Carlo study for a table of percentage points gives a single value $\hat{X}_p$ which is the "best" estimate of ξ_p, the question arises as to how this should be done. Is it best to make c runs and then take the mean $\bar{X}_p$ or make one large run, then take X_p as the point estimate? In both cases the point estimates are biased in finite samples. David (1970, page 65) gives expansions for the expected value and variance of the k^{th} order statistic in a sample of size n in terms of the inverse cumulative distribution function, $Q(x)$, of the parent population and its derivatives.

Let $p_k = k/(n+1)$, $q_k = 1-p_k$, and let $Q^{(i)}(x)$ be the i^{th} derivative of $Q(\cdot)$ evaluated at x. Then

$$E(X_{(k)}) = Q(p_k) + \frac{p_k q_k}{2(n+2)} Q^{(2)}(p_k) + \cdots . \tag{3.1}$$

For $X_{(k)}$, with $k = [np]+1$, to be an unbiased estimator of ξ_p, we need $E(X_{(k)}) = \xi_p = Q(p)$. For certain values of n and p, $Q(p_k)$ will equal $Q(p)$, and the bias in $X_{(k)}$ is given by the second and subsequent terms in (3.1). This would be so, for example if $n = 99$, $p = .8$; then $k = [np]+1 = 80$ and the first term of (3.1) is $Q(80/100) = Q(.8) = \xi_{.8}$ as required. However, in general this will not be so, and to investigate the bias we expand $Q(p_k)$ about $Q(p)$ and $Q^{(2)}(p_k)$ about $Q^{(2)}(p)$, using Taylor's series. Then (3.1) becomes

$$E(X_{(k)}) = Q(p) + (p_k-p)Q^{(1)}(p) + (p_k-p)^2Q^{(2)}(p)/2 \ldots$$

$$+\frac{p_kq_k}{2(n+2)}\,(Q^{(2)}(p) + (p_k-p)Q^{(3)}(p) \ldots + \ldots \ .$$

Since $p_k-p = \frac{[np]+1}{n+1} - p = \frac{1-p-\varepsilon}{n+1}$, where $0 \leq \varepsilon \leq 1$, we have, to terms of order $1/n$,

$$(3.2) \qquad E(X_{(k)}) = Q(p) + \frac{1-p-\varepsilon}{n+1}\,Q^{(1)}(p) + \frac{p(1-p)}{2(n+2)}\,Q^{(2)}(p) \ .$$

When several samples are used, the estimate of ξ_p is $\bar{X}_{(\ell)}$, the mean of the c values of $X_{(\ell)}$, where ℓ is $[mp]+1$. Then

$$E(\bar{X}_{(\ell)}) = E(X_{(\ell)}) = Q(p) + \frac{1-p-\varepsilon}{m+1}\,Q^{(1)}(p) + \frac{p(1-p)Q^{(2)}(p)}{2(m+2)}$$

to order $1/m$, where again $0 < \varepsilon < 1$.

The bias in using the single-sample estimate is therefore reduced by a factor of approximately c.

3.2. Variance of estimated percentage points, and length of confidence intervals.

Also from David (1970, page 65) we have, for the single sample of size n:

$$\mathrm{Var}(X_{(k)}) = \frac{p_kq_k}{n+2}\,\{Q^{(1)}(p_k)\}^2 + \text{terms in } 1/(n+2)^2 \ .$$

Thus to order $1/n$,

$$\mathrm{Var}(X_{(k)}) = \frac{p(1-p)}{n+2} \{Q^{(1)}(p)\}^2 \text{ , and the variance of } \bar{X}_{(\ell)} \text{ is}$$

$$\mathrm{Var}(\bar{X}_{(\ell)}) = \frac{p(1-p)}{c(m+2)} \{Q^{(1)}(p)\}^2$$

to order $1/m$. For large samples the ratio of these variances is 1. The expected length of the confidence interval for ξ_p, given by $\bar{X}_p$ will be approximately, to order $n^{-1/2}$,

$$L_M = 2z_{1-\alpha/2} \{\frac{p(1-p)}{n}\}^{1/2} Q^{(1)}(p) \; ; \tag{3.3}$$

the expected length of $L \equiv (X_{(r)}, X_{(s)})$, using (3.1) with (2.1) and (2.2) will be approximately

$$L_S = 2z_{1-\alpha/2} (\frac{p(1-p)}{n})^{1/2} Q^{(1)}(p) \; .$$

Thus the lengths of the confidence intervals given by the two methods can be expected to be roughly the same size, diminishing as $n^{-1/2}$. There is of course a differnce in the nature of the confidence intervals: the interval obtained by the multisample method will have endpoints equally spaced about the estimate $\bar{X}_p$, while L is designed so that the endpoints have estimated significance levels equally spaced about p. For many uses of confidence intervals the second property may well be the more desirable.

3.3. Monte Carlo Results.

Some Monte Carlo studies have been made which illustrate the above points. Samples of size n were taken from a standard normal distribution and each sample was used to estimate the 95% and 97.5% percentage point; the sample was then randomly divided into $c = 10$ groups, each with $m = n/c$ observations, and the estimates found by $\bar{X}_p$ as described above. Results are given in Table 1. Since the same set of n values was used for both methods, the resulting estimates are correlated; nevertheless the larger bias in the multi-sample estimate can be seen for small n. When n increases, each subsample becomes sufficiently large that the individual estimates will become quite accurate and $\bar{X}_p$ loses its bias. The approximate expected length of the confidence interval, given by (3.3) is given in the Table. The approximation gives good results for $p = 0.95$, but the estimate appears to be too large for $p = 0.975$. Nevertheless, it can be seen that there is little to choose between the length of the intervals given by the two methods. Table 2 shows the actual probability levels corresponding to the estimates $\hat{\xi}_p$ given by both methods: again, the smaller bias for the single sample method is shown, for relatively small n. It is this smaller bias which would appear to give an advantage to the single sample method, since point estimates and not confidence intervals are usually what are required for Monte Carlo percentage points.

This research was supported in part by the National Research Council of Canada (now the National Science and Engineering Research Council), and by the U.S. Office of Naval Research, Contract N00014-76-C-0475, and the ıthors thank these agencies for their support.

References

David, H.A. (1970). Order Statistics. John Wiley; New York.

Juritz, J.M., Juritz, J.W.F., and Stephens, M.A. (1978). The use of order statistics to assess the precision of simulations. Unpublished manuscript, Statistics Department, Stanford University.

Schafer, R.E. (1974). On assessing the precision of simulations. J. Statist. Compt. Simul., 3, 67-69.

TABLE 1. ESTIMATES OF PERCENTILES OF A STANDARD NORMAL DISTRIBUTION.

The table gives estimates of ξ_p, the p-th percentile, upper and lower 95% confidence limits, and the length of the confidence interval. For the single method n variables were used, and for the multi-sample method the same variables were used in ten groups; m = n/10.

p = 0.95 True value of ξ_p = 1.6449.

Method	m	n	Estim. point ξ_p	Lower limit	Upper limit	Length	Difference in length single-multi	Approx. expected length
Multi-sample	50		1.726	1.577	1.875	.298		
Single-sample		500	1.643	1.520	1.785	.265	-.033	.370
Multi-sample	100		1.685	1.566	1.804	.238		
Single-sample		1000	1.665	1.570	1.775	.205	-.033	.262
Multi-sample	200		1.688	1.595	1.781	.186		
Single-sample		2000	1.677	1.620	1.740	.120	-.066	.185
Multi-sample	500		1.671	1.634	1.708	.074		
Single-sample		5000	1.670	1.633	1.715	.082	.008	.117
Multi-sample	1000		1.648	1.610	1.686	.076		
Single-sample		10000	1.655	1.625	1.680	.055	-.021	.083
p = 0.975 True value of ξ_p = 1.9600.								
Multi-sample	50		1.994	1.745	2.243	.498		
Single-sample		500	1.970	1.824	2.160	.336	-.162	.468
Multi-sample	100		1.986	1.876	2.096	.220		
Single-sample		1000	1.965	1.875	2.110	.235	.015	.331
Multi-sample	200		1.972	1.865	2.079	.214		
Single-sample		2000	2.016	1.927	2.078	.151	-.063	.234
Multi-sample	500		2.003	1.946	2.060	.114		
Single-sample		5000	1.998	1.970	2.055	.115	.001	.148
Multi-sample	1000		1.944	1.926	1.962	.036		
Single-sample		10000	1.943	1.909	1.976	.067	.031	.105

TABLE 2.

Actual probability levels of the estimated percentile X_p or $\overline{X}_p$ for a percentile ξ_p of the N(0,1) distribution.

n	Method	
	Multi-sample	Single-sample
	p = 0.975	
500	.9769	.9756
1000	.9765	.9753
2000	.9757	.9786
5000	.9774	.9771
10000	.9740	.9740
	p = 0.95	
500	.9578	.9499
1000	.9540	.9520
2000	.9543	.9532
5000	.9526	.9525
10000	.9503	.9510

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 309	2. GOVT ACCESSION NO. AD-A109 660	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON THE ACCURACY OF SIMULATED PERCENTAGE POINTS		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) J.M. Juritz, J.W.F. Juritz, and M.A. Stephens		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0475
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office Of Naval Research Statistics & Probability Program Code 436 Arlington, VA 22217		12. REPORT DATE OCTOBER 7, 1981
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Monte Carlo methods; Percentage points; Significance points; Simulated percentage points; Simulations.

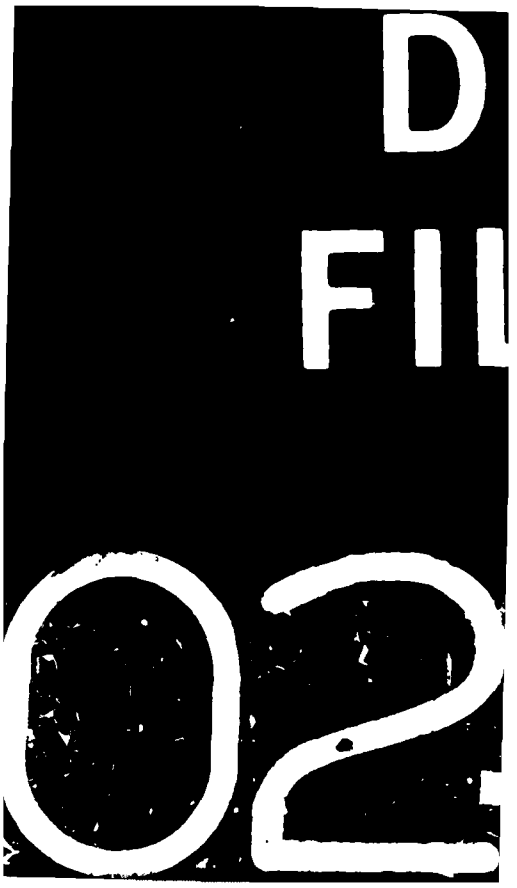
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The estimates of percentage points of a distribution, obtained from one large sample of Monte Carlo values, are compared with those obtained by dividing the large sample into several subsamples and taking the average. The single-sample method appears to be preferable.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


D
FI
02